Express Mail No. EV 222905732 US

METHOD AND SYSTEM FOR PREVENTING DEADLOCK IN FIBRE CHANNEL FABRICS USING FRAME PRIORITIES

INVENTOR:

STEVEN M. BETKER

Cross Reference to related Applications

[0001] This application claims priority to U.S. provisional patent application serial number 60/542,186, filed on 02/05/2004, the disclosure of which is incorporated herein by reference in its entirety.

BACKGROUND

Field of the Invention

[0002] The present invention relates to Fibre Channel systems, and more particularly, to reducing deadlock problems in Fibre Channel Fabrics.

Background of the Invention

[0003] Fibre channel is a set of American National Standard Institute (ANSI) standards, which provide a serial transmission protocol for storage and network protocols such as HIPPI, SCSI, IP, ATM and others. Fibre channel provides an input/output interface to meet the requirements of both channel and network users.

[0004] Fibre channel supports three different topologies: point-to-point, arbitrated loop and fibre

channel fabric. The point-to-point topology attaches two devices directly. The arbitrated loop topology attaches devices in a loop. The fibre channel fabric topology attaches host systems directly to a fabric, which are then connected to multiple devices. The fibre channel fabric topology allows several media types to be interconnected.

[0005] Fibre channel is a closed system that relies on multiple ports to exchange information on attributes and characteristics to determine if the ports can operate together. If the ports can work together, they define the criteria under which they communicate.

[0006] In fibre channel, a path is established between two nodes where the path's primary task is to transport data from one point to another at high speed with low latency, performing only simple error detection in hardware.

[0007] Fibre channel fabric devices include a node port or "N_Port" that manages fabric connections. The N_port establishes a connection to a fabric element (e.g., a switch) having a fabric port or F_port. Fabric elements include the intelligence to handle routing, error detection, recovery, and similar management functions.

[0008] A fibre channel switch is a multi-port device where each port manages a simple point-to-point connection between itself and its attached system. Each port can be attached to a server, peripheral, I/O subsystem, bridge, hub, router, or even another switch. A switch receives messages from one port and automatically routes it to another port. Multiple calls or data transfers happen concurrently through the multi-port fibre channel switch.

[0009] Fibre channel switches use memory buffers to hold frames received and sent across a network. Associated with these buffers are credits, which are the number of frames that a buffer can hold per fabric port.

[0010] The following Fibre Channel standards are used for Fibre Channel systems and Fibre Channel Fabrics, and are incorporated herein by reference in their entirety:

[0011] ANSI INCITS xxx-200x Fibre Channel Framing and Signaling Interface (FC-FS) - T11/Project 1331D; and ANSI INCITS xxx-200x Fibre Channel Switch Fabric-3 (FC-SW-3), T11/Project 1508D.

[0012] As discussed above, a Fibre Channel Fabric can consist of multiple switches connected in an arbitrary topology. The links between the switches use a

buffer-to-buffer credit scheme for flow control so that all frames transmitted have a receive buffer. Fabric deadlock may occur if a switch cannot forward frames because the recipient switch buffers (receive buffers) are full.

[0013] The following example, described with respect to Figure 1E, shows how a deadlock situation can occur. Figure 1E shows five switches ("SW") 1, 2, 3, 4, and 5 that are linked together by ISLs (Inter Switch Links) in a ring topology. Host 11 and target 21 are linked to switch 1, host 12 and target 22 are linked to switch 2, and so forth.

[0014] In this example, hosts 11-15 can send data as fast as they can to a target that is two (2) hops (number of ISLs) away, for example:

Host 11 can send data to target 23;

Host 12 can send data to target 24;

Host 13 can send data to target 25;

Host 14 can send data to target 21; and

Host 15 can send data to target 22

[0015] For illustration purposes only, all traffic goes in the clockwise direction in Figure 1E.

[0016] The receive buffers available for each ISL in the direction of traffic may get filled with frames addressed to the next switch. For example:

For the ISL between switch 1 and switch 2, the receive buffers on switch 2 get filled with frames for switch 3;

For the ISL between switch 2 and switch 3, the receive buffers on switch 3 get filled with frames for switch 4;

For the ISL between switch 3 and switch 4, the receive buffers on switch 4 get filled with frames for switch 5;

For the ISL between switch 4 and 5, the receive buffers on 5 get filled with frames for switch 1; and

For the ISL between switch 5 and switch 1,the receive buffers on switch 1 get filled with frames for switch 2.

[0017] The transmit side of a switch waits for R_RDYs before it can transmit any frames. If frames cannot be transmitted from one ISL, then the receive buffers for the other ISL cannot be emptied. If the receive buffers cannot be emptied, no R_RDY flow control signals can be transmitted, which deadlocks the Fabric.

[0018] Many large Fabrics have paths that form rings within them, especially if they are designed to avoid single points of failure by using redundant switches. Such network traffic patterns may result in a deadlock situation disrupting networks using fibre channel switches and components.

[0019] Therefore, there is need for a system and method for minimizing deadlock problems in fibre channel switches.

SUMMARY OF THE PRESENT INVENTION

[0020] In one aspect of the present invention, a method for transmitting frames using a fibre channel switch is provided. The method includes, determining a frame's priority based on a hop count for the frame; placing a frame in a priority queue, where the priority queue is dedicated to frames having similar priorities; selecting a frame for transmission based on the frame's priority, if credit is available, where a frame with a higher priority is sent before a frame with a lower priority; and selecting a frame with a lower priority if enough higher priority frames have been sent.

[0021] In another aspect of the present invention, a system for transmitting fibre channel frames is provided. The system includes a switch with at least two priority queues for placing frames with different priorities, where a frame's priority is based on a hop count depending upon the frame's destination; a counter that keeps track of frames that are transmitted from the two priority queues; and a credit control module that determines if credit is available before sending a particular frame.

[0022] In yet another aspect of the present invention, a fibre channel switch having receive and transmit ports for transmitting frames is provided.

The switch includes, at least two priority queues for placing frames with different priorities, where a frame's priority is based on a hop count depending upon the frame's destination; a counter that keeps track of frames that are transmitted from the two priority queues; and a credit control module that determines if credit is available before sending a particular frame.

[0023] In yet another aspect, a system for transmitting fibre channel frames is provided. The system includes, means for placing a frame in a priority queue, where the priority queue is dedicated to frames having similar priorities; means for selecting a frame for transmission based on the frame's priority, if credit is available, where a frame with a higher priority is sent before a frame with a lower priority; and means for selecting a frame with a lower priority if enough higher priority frames have been sent.

[0024] In yet another aspect of the present invention, a fibre channel switch having a receive port and a transmit port for transmitting fibre channel frames is provided. The switch includes, means for placing a frame in a priority queue, where the priority queue is dedicated to frames having similar priorities; means for selecting a frame for transmission based on

the frame's priority, if credit is available, where a frame with a higher priority is sent before a frame with a lower priority; and means for selecting a frame with a lower priority if enough higher priority frames have been sent.

[0025] This brief summary has been provided so that the nature of the invention may be understood quickly. A more complete understanding of the invention can be obtained by reference to the following detailed description of the preferred embodiments thereof in connection with the attached drawings.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0026] Definitions:

[0027] The following definitions are provided as they are typically (but not exclusively) used in the fibre channel environment, implementing the various adaptive aspects of the present invention.

[0028] "D_ID": 24-bit fibre channel header field that contains destination address.

[0029] "EOF": End of Frame

[0030] "E-Port": A fabric expansion port that attaches to another Interconnect port to create an Inter-Switch Link.

[0031] "F-Port": A port to which non-loop N_Ports are attached to a fabric and does not include FL_ports.

[0032] "Fibre channel ANSI Standard": The standard (incorporated herein by reference in its entirety) describes the physical interface, transmission and signaling protocol of a high performance serial link for support of other high level protocols associated with IPI, SCSI, IP, ATM and others.

[0033] "FC-1": Fibre channel transmission protocol, which includes serial encoding, decoding and error control.

[0034] "FC-2": Fibre channel signaling protocol that includes frame structure and byte sequences.

[0035] "FC-3": Defines a set of fibre channel services that are common across plural ports of a node.

[0036] "FC-4": Provides mapping between lower levels of fibre channel, IPI and SCSI command sets, HIPPI data framing, IP and other upper level protocols.

[0037] "Fabric": The structure or organization of a group of switches, target and host devices (NL_Port, N_ports etc.).

[0038] "Fabric Topology": A topology where a device is directly attached to a fibre channel fabric that uses destination identifiers embedded in frame headers to route frames through a fibre channel fabric to a desired destination.

[0039] "FL_Port": A L_Port that is able to perform the function of a F_Port, attached via a link to one or more NL_Ports in an Arbitrated Loop topology.

[0040] "Inter-Switch Link" ("ISL"): A link directly connecting the E_port of one switch to the E_port of another switch.

[0041] Port: A general reference to N. Sub.__ Port or F.Sub.__Port.

[0042] "L_Port": A port that contains Arbitrated Loop functions associated with the Arbitrated Loop topology.

[0043] "N-Port": A direct fabric attached port.

[0044] "NL_Port": A L_Port that can perform the function of a N_Port.

[0045] "R_RDY": Flow control primitive signal used for establishing credit. Receiving an R_RDY frame increases credit, while sending a R_RDY frame decreases credit.

[0046] "S_ID": 24-bit fibre channel header field that contains the source address of a frame.

[0047] "SOF": Start of Frame

[0048] "Switch": A fabric element conforming to the Fibre Channel Switch standards.

[0049] To facilitate an understanding of the preferred embodiment, the general architecture and

operation of a fibre channel system will be described. The specific architecture and operation of the preferred embodiment will then be described with reference to the general architecture of the fibre channel system.

[0050] Figure 1A is a block diagram of a fibre channel system 100 implementing the methods and systems in accordance with the adaptive aspects of the present invention. System 100 includes plural devices that are interconnected. Each device includes one or more ports, classified as node ports (N_Ports), fabric ports (F_Ports), and expansion ports (E_Ports). Node ports may be located in a node device, e.g. server 103, disk array 105 and storage device 104. Fabric ports are located in fabric devices such as switch 101 and 102. Arbitrated loop 106 may be operationally coupled to switch 101 using arbitrated loop ports (FL_Ports).

[0051] The devices of Figure 1A are operationally coupled via "links" or "paths". A path may be established between two N_ports, e.g. between server 103 and storage 104. A packet-switched path may be established using multiple links, e.g. an N-Port in server 103 may establish a path with disk array 105 through switch 102.

[0052] Figure 1B is a block diagram of a 20-port ASIC fabric element according to one aspect of the present invention. Figure 1B provides the general architecture of a 20-channel switch chasis using the 20-port fabric element. Fabric element includes ASIC 20 with non-blocking fibre channel class 2 (connectionless, acknowledged) and class 3 (connectionless, unacknowledged) service between any ports. It is noteworthy that ASIC 20 may also be designed for other fibre channel classes of service, within the scope and operation of the present invention as described herein.

[0053] The fabric element of the present invention is presently implemented as a single CMOS ASIC, and for this reason the term "fabric element" and ASIC are used interchangeably to refer to the preferred embodiments in this specification. Although Figure 1B shows 20 ports, the present invention is not limited to any particular number of ports.

[0054] ASIC 20 has 20 ports numbered in Figure 1B as GL0 through GL19. These ports are generic to common Fibre Channel port types, for example, F_Port, FL_Port and E-Port. In other words, depending upon what it is attached to, each GL_Port can function as any type of

port. Also, the GL_Port may function as a special port useful in fabric element linking, as described below.

[0055] For illustration purposes only, all GL_Ports are drawn on the same side of ASIC 20 in Figure 1B. However, the ports may be located on both sides of ASIC 20 as shown in other figures. This does not imply any difference in port or ASIC design. Actual physical layout of the ports will depend on the physical layout of the ASIC.

[0056] Each port GL0-GL19 has transmit and receive connections to switch crossbar 50. One connection is through receive buffer 52, which functions to receive and temporarily hold a frame during a routing operation. The other connection is through a transmit buffer 54.

[0057] Switch crossbar 50 includes a number of switch crossbars for handling specific types of data and data flow control information. For illustration purposes only, switch crossbar 50 is shown as a single crossbar. Switch crossbar 50 is a connectionless crossbar (packet switch) of known conventional design, sized to connect 21 x 21 paths. This is to accommodate 20 GL_Ports plus a port for connection to a fabric controller, which may be external or internal to ASIC 20.

[0058] In the preferred embodiments of switch chasis described herein, the fabric controller is a firmware-programmed microprocessor, also referred to as the input/output processor "IOP"). IOP 66 is shown in Figure 1C as a part of a switch chasis utilizing one or more of ASIC 20. As seen in Figure 1B, bi-directional connection to IOP 66 is routed through path 67, which connects internally to a control bus 60. Transmit buffer 56, receive buffer 58, control register 62 and Status register 64 connect to bus 60. Transmit buffer 56 and receive buffer 58 connect the internal connectionless switch crossbar 50 to IOP 66 so that it can source or sink frames.

[0059] Control register 62 receives and holds control information from IOP 66, so that IOP 66 can change characteristics or operating configuration of ASIC 20 by placing certain control words in register 62. IOP 66 can read status of ASIC 20 by monitoring various codes that are placed in status register 64 by monitoring circuits (not shown).

[0060] Figure 1C shows a 20-channel switch chassis S2 using ASIC 20 and IOP 66. S2 will also include other elements, for example, a power supply (not shown). The 20 GL_Ports correspond to channel C0-C19. Each GL_Port has a serial/deserializer (SERDES)

designated as S0-S19. Ideally, the SERDES functions are implemented on ASIC 20 for efficiency, but may alternatively be external to each GL_Port.

[0061] Each GL_Port may have an optical-electric converter, designated as OE0-OE19 connected with its SERDES through serial lines, for providing fibre optic input/output connections, as is well known in the high performance switch design. The converters connect to switch channels C0-C19. It is noteworthy that the ports can connect through copper paths or other means instead of optical-electric converters.

[0062] Figure 1D shows a block diagram of ASIC 20 with sixteen GL_Ports designated as GL0-GL15 and four 10G port control modules designated as XG0-XG3. ASIC 20 include a control port 62A that is coupled to IOP 66 through a PCI connection 66A.

[0063] In the preferred embodiments of switch chassis described herein, the switch controller is a firmware-programmed microprocessor (IOP 66). IOP 66 is also shown in Figure 2 as a part of a switch chassis 201, containing switch ports 204, 207, 210 and 215. Each port as described above has a transmit port (segment), for example, 205, 208, 211 and 213, and receive port (segment), for example, 206, 209, 212 and

214, that have been described above with respect to Figures 1B-1D.

[0064] Transmit and receive ports are connected by switch crossbar 50 so that they can transfer frames. IOP 66 controls and configures the switch ports.

[0065] In one aspect, the present invention prevents deadlocks on E-Ports by placing frames queued for transmission at a transmit port (for example, 205 in Figure 2). The frames are placed in separate queues based on the number of "switch to switch" hops to a destination. Frames with lower hop counts get higher priority over frames that have higher hop counts. A system and method is provided such that low priority frames are also transmitted (especially when enough high priority frames cannot be transmitted). This allows all frames to be transmitted in a finite amount of time regardless of load, as long as the frame destinations can accept frames within a finite amount of time (which means all frames that enter the switch are delivered to their destination N-Ports). The number of receive buffers (for example, 206) for receiving incoming frames is greater than or equal to the maximum number of hops to destination switches.

[0066] In one aspect of the present invention, Figure 3 shows a transmit port (e.g. 205) with a

transmit frame (port) queue module 301 (also referred to herein as "module 301") and buffer-to-buffer credit module 302. Frames are queued in module 301 as they are received from other ports (including receive ports, for example, 206) and routed to the transmit port (in this example, 205). Frames are transmitted (301B) from module 301 to a device linked to that port (e.g., 404, Figure 4).

[0067] The buffer-to-buffer credit module 302 ensures that frames are only sent if the receiving end (i.e. the device/port that receives frame 301B(not shown in this example)) has a buffer available to receive the transmitted frame. Buffer to buffer credit module receives R_RDYs 302A from a receive port (in this example, 206). As described in FC-FS and FC-SW-3 (incorporated herein by reference in its entirety), a buffer-to-buffer credit count is initialized during port login. The count is decremented whenever a frame is sent. The count is incremented whenever an R_RDY primitive is received from the other end of the link.

[0068] Figure 4 shows transmit queue module 301 used to implement transmission priorities and decide which frame to transmit to port 404 of another switch (not shown). Port 404 is a port on another switch that is

connected to transmit module 403 by a standard Fibre Channel cable.

[0069] Frame priority corresponds to a hop count, where hop count is the number of ISLs a frame has to traverse before it gets to its destination. For example, in Figure 1E, for a frame being sent by Switch 1 to Switch 3 has a hop count of 2. The hop count for each destination is derived from the standard FSPF routing data exchanged by switches as described in FC-SW-3 standard.

[0070] In this embodiment each frame queued for transmission at a transmit port (for example, port 205) is assigned a priority number that is one less than the hop count, for example, a frame having a hop count of 2 has a priority 1. In this example, a lower priority number means that the frame has higher priority. However, the invention is not limited to how the priority numbers are assigned, for example, a higher number may be assigned to higher priority frames, as long as the hop count is used to assign the priority, and lower hop counts have higher priority.

[0071] Module 301 has an individual queue for each priority number. A frame is placed in a particular queue based on its priority number. For example, queue 401 has N-1 queues that are used for placing frames.

Priority queue 0 keeps frames with priority number 0 (in this example, the highest priority frame), priority queue 1 keeps frames that have priority number 1, and so forth.

[0072] Every priority queue (0 to N-1) has a counter 402 that is used to avoid a situation where low priority frames are not sent because a switch has a constant flow of higher priority frames. Each counter 402 is initialized to 0 when transmit port 205 is initialized. Transmit module 403 uses the method described below to select a queue for frame transmission.

[0073] As described above, a frame with a lower hop count gets priority over a frame with a higher hop count. Each of counters 402 counts the number of tries that are made by the queue(s) to transmit a frame. A lower priority frame can be sent if the counter for the next highest priority is at 2 or if there are no higher priority frames and the total transmit credit available is greater than what is needed for the lower priority frame.

[0074] The number "2" ensures that more higher priority frames than lower priorities are sent if the higher priority frames are queued. A lower priority frame is not sent until either the sum of the empty

receive buffers at port 404, and receive buffers filled with higher priority frames at 404, is greater than or equal the hop count for the lower priority frame. This ensures that higher priority frames can always be sent even after lower priority frames are sent, because enough receive buffers in 404 are either empty or contain higher priority frames that will be able to move on and empty those buffers.

[0075] It is noteworthy that the present invention is not limited to a counter value "2", any other value may be used to adapt the aspects of the present invention.

[0076] The foregoing allows frames with lower priority to be transmitted, while ensuring that whenever a lower priority frame is sent, the number of receive buffers at port 404 that are either empty or contain higher priority frames is greater than or equal to the hop count of the frame just transmitted. Counter 402 is cleared to zero whenever a lower priority frame (i.e. in this example, with a higher priority number) is sent.

[0077] The following provides an example with respect to Figure 1E. The transmit port on switch 1 that is connected through the ISL to switch 2 has frames with hop count 1 (received from switch 5,

destination target 22) and frames with hop count 2 (received from host 11, destination target 23). For every 2 frames sent with 1 hop count, 1 frame with hop count 2 is sent. Whenever a frame with hop count 2 is sent, at least one of the receive buffers on switch 2 is either empty or has a frame with destination target 22, which it can send and then empty the buffer. So frames with hop count 1 can always be sent, and the frames with hop count 2 can be sent after waiting for enough hop count 1 frames to be sent.

[0078] Figure 5 shows a flow diagram of how frames are selected for transmission, according to one aspect of the present invention. It is noteworthy that combinatorial hardware logic may be used to select frames in a single clock cycle, according to one aspect of the present invention. The process starts in step S501 when a transmit port (in this example, 205) is ready to send a frame.

[0079] In step S502, if no transmit queues (401) have any frames, then port 205 waits for frames. If there are queued frames in 401, the process moves to step S503.

[0080] In step S503, the process selects the highest priority transmit queue (i.e. in this example, the

queue with the lowest priority number (0 to N-1)) that has frames queued for transmission.

[0081] In step S504, the process determines if credit is available for transmitting the frame from the particular priority queue. This is performed by buffer-to-buffer credit module 302(Figure 3). If the available credit is less than what is required for the frame, then the process goes back to step S502. It is noteworthy that the available credit may change if R_RDY primitives are received later. If available credit is greater than or equal to what is required for the frame, the process goes to step S505.

[0082] In step S505, the process compares the count associated with a particular queue. Counter 402 performs this. In one aspect the count is compared with 2.

[0083] If the count is less than 2, the process goes to step S506. In step S506, the queue count (by counter 402) is incremented by 1, and the process goes to step S507. In step S507, if the particular queue is empty, the process goes back to S501, otherwise the frame is sent in step S508.

[0084] If the count is greater than 2, then in step S509 the count for the queue is cleared to 0, and the process goes to 510. In step S510, the process

determines if this is the last (lowest priority) queue, if yes, then the process goes back to step S501. Otherwise the process moves to step S511. In step S511, the next highest priority queue is selected. This provides lower priority queues a chance to send frames. The process moves to step S505 to see if a frame from that queue can be transmitted.

[0085] It is assumed that all frames that arrive at a destination switch are delivered to N-ports. This means that all frames sent by a switch with priority 0 (1 hop count) will be delivered and the receive buffers at the receive end of the ISL will be freed, with R_RDYs being sent. Since priority 0 is highest, at least one receive buffer on every ISL will either be filled with a frame sent as priority 0, or be empty. Hence all frames queued at priority 0 can be sent, and all the receive buffers used for them can be cleared.

[0086] If all frames in a switch queued at priority N or higher can be sent, and if at any time there are at least N+1 receive buffers with higher priority frames or empty, then all neighboring switches will always be able to send priority N+1 frames (which become priority N when received). The requirement of at least N+1 receive buffers that are either empty or filled with higher priority frames is ensured by the

algorithm for sending lower priority frames, which only sends lower priority frames if at least 2**N higher priority frames have been sent, or if transmit credit (empty receive buffers) is greater than N.

[0087] In one aspect of the present invention, lower priority frames can be sent and deadlock situations can be reduced. This improves the overall efficiency of a network using fibre channel switches.

[0088] Although the present invention has been described with reference to specific embodiments, these embodiments are illustrative only and not limiting. Many other applications and embodiments of the present invention will be apparent in light of this disclosure and the following claims.